# Bridging the Communication Gap between ML Engineers and Data Enrichment Workers

**Claudel Rheault, Jerome Pasquero, Patrick Steeves, Karina Pannhasith**

Sama
crheault, jpasquero, psteeves, kpannhasith@sama.com

**Abstract**

Human-annotated and -validated data remains a core component of machine learning (ML) model development. Yet, a significant portion of the expertise gathered by data enrichment workers goes undocumented, largely because the appropriate tools are lacking and there's no platform for their voices to be acknowledged. We introduce a series of prototypes that allow these workers to convey their findings, ensuring they are engaged in a dialogue with ML developers, rather than being overlooked.

## Introduction

The development of modern computer vision models heavily relies on human-labeled data (Lin et. al. 2014; Sun et. al. 2020; Kuznetsova et. al. 2020). Historically, machine learning algorithms have feasted on vast amounts of training data annotated entirely from scratch. Models like ResNet, YOLO, and R-CNN and VGG derivatives all raised the state of the art by relying on manually annotated datasets (Geirhos et. al. 2021; Emam et. al. 2021).

Given that the performance of newer (and bigger) models keeps scaling with the volume of training data, the industry is focusing on reducing the reliance of these models on manual labeling. This is understandable, as the cost of labeling billions of data points becomes prohibitive for even the best funded labs and enterprises. In this endeavor, foundation models and synthetic data generation are beginning to show promise. The former can significantly speed up the labeling process by automating most of it (Schilling et. al. 2021; Fernández-Moreno et. al. 2023), while the latter can generate entirely new scenes already labeled (Borkman et. al. 2021; Wood et. al. 2021).

The potential of these nascent technologies might suggest an imminent reduction in the importance of data enrichment work in the model development lifecycle. However, while manual annotation may play a lesser role, data enrichment as a whole will become more critical than ever, especially as models make it out of the lab and into production. Successful use cases for AI will leverage the deep understanding of data workers throughout the training, evaluation, and monitoring phases. Programmatic or automatic approaches are helpful tools, but human involvement is essential to ensure performant, robust, and safe deployment. Edge cases, bias, and domain shifts all fall outside of a model's learnable domain and require human judgment to properly identify and understand. In other words, models cannot monitor themselves. Unfortunately, most of the industry considers data enrichment work to be simple, repetitive, and unworthy of much attention. In most organizations, we witness a dangerous chasm between those closest to the data and those developing the models.

## Motivation for the work

We find ourselves at a pivotal moment in the history of computer vision data enrichment work where we have the opportunity to further elevate the contribution of enrichment workers to model development and monitoring cycles. The current risk is that the gap between workers and developers will grow. In this scenario, workers will receive less and less context surrounding the annotation or validation work they have to perform, adding to their cognitive load and disengagement, and reducing their performance (Dai et. al. 2015; Kaufmann et. al. 2011; Kittur et. al. 2013). On the other hand, developers will forgo valuable and often subtle insights that workers accumulate by spending time with the data, leading them to overlook potentially critical model failure modes.

We propose harnessing data enrichment workers' expertise and intuition honed through careful examination of data and correction of model predictions. Equipped with the right tools and training, data enrichment workers can communicate indispensable insights from the data to guide machine learning (ML) engineering in improving model performance and robustness. Our objective is to collaborate with data enrichment workers in co-designing and developing a set of tools and interactions that empower them to enrich data while extracting informative insights beyond what existing models can achieve on their own. Envisioning a future where data enrichment workers

harness the scaling powers of state-of-the-art machine learning technology, we hope to increase the amount of computer vision data they review and the quantity and quality of analyses they extract from it. Insights they find could include systematic model failure modes (e.g., mispredictions of stop signs at sunset), edge cases (e.g., stop signs with graffiti), sensor issues (e.g. poorly calibrated 2D and 3D sensors on a vehicle), or any other scenario that automated analysis cannot capture. To achieve success, we emphasize that the tools we aim to develop should foster a strong partnership with the ML Engineering teams responsible for model deployment and maintenance. Therefore, the concepts we present here are primarily focused on incorporating robust collaboration features. Our vision is to create an ecosystem where seamless communication and knowledge exchange between all stakeholders contribute to the continuous improvement and efficiency of the model development and production processes.

## Workers' perspectives

We surveyed a sample of data enrichment workers to understand their perspective on what value they feel they could provide in the data handling process (apart from their current work annotating and correcting annotations). We also wanted to understand the level of context they want about the AI application they are helping build. A total of 70 respondents participated who work on 2D and 3D automotive and agricultural data. When asked how much context they'd prefer to have in order to do good annotations, 84.3% responded they want high context (HIGH - I want as much context on the model as I can - what is the AI tool being built by the client), as compared to medium and low context.

They were subsequently asked to detail what they would do with the extra context if it was given. Some workers mentioned that they would like to understand their broader impact: *"Get to understand the why. i.e. why we annotating"* or to *"Motivate me to know I am helping clients to achieve their goals, so that when I see the end product on the market I become proud knowing I took part in it."* Some others see the context as something that would enable them to give proper feedback on the work: *"it will help me work better and be able to provide sufficient feedback"* or *"getting to know more about the context and share my knowledge on what is working and not working."* The survey then inquired about 6 other types of data work tasks they could be involved in, and they had to indicate the ones where they feel they could add most value (1) to least value (6). The most value respondents felt they could add was in the task "Highlight the most important things you've seen in the data." This exploratory survey gives us confidence that tools to encourage workers in other types of data work would be well received by our teams.

## Tools for the data work of tomorrow

We introduce two prototypes that encourage the sharing of insights gained from hands-on interaction with data. The prototypes aim to foster a rich, mutually beneficial connection between ML experts responsible for developing and refining models and the data enrichment specialists dedicated to generating essential training and validation data. Previous work has explored instruction building with users (Manam and Quinn, 2018), human in the loop data discovery (Han, Dong, and Demartini, 2021) and tools for workers to collaborate within themselves (Gray et. al. 2016; Irani and Silberman 2013). Our prototypes intend to facilitate information exchange and empower data workers to develop new types of skills in data understanding.

### Calibration Log

More often than not, data annotation platforms are catered towards the user experience needs of the ML engineering team, rather than the needs of the data enrichment workers, which overlooks their significant contributions. Nonetheless, the importance of open dialogue between these groups has been recognized, and the active involvement of data professionals in shaping annotation instructions has proven invaluable in not only elevating annotation quality but also enhancing the worker's experience (Partnership on AI, 2021; Miceli, Schuessler, and Yang, 2020).

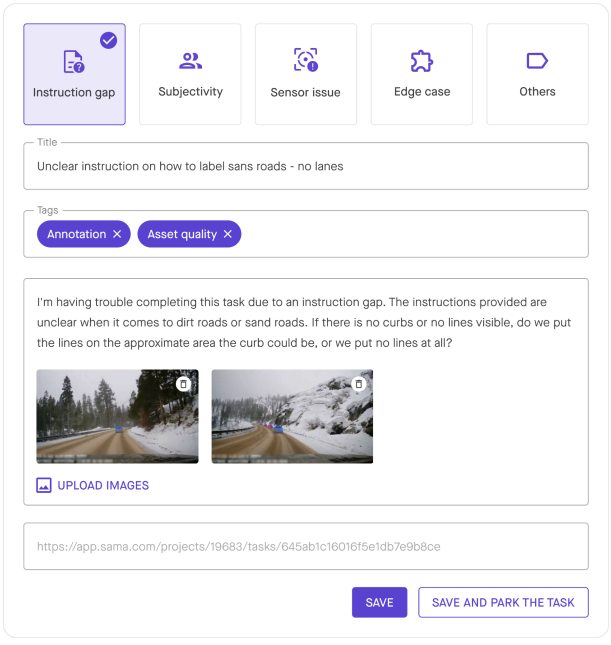


Figure 1: Calibration Log - main feedback dialog box

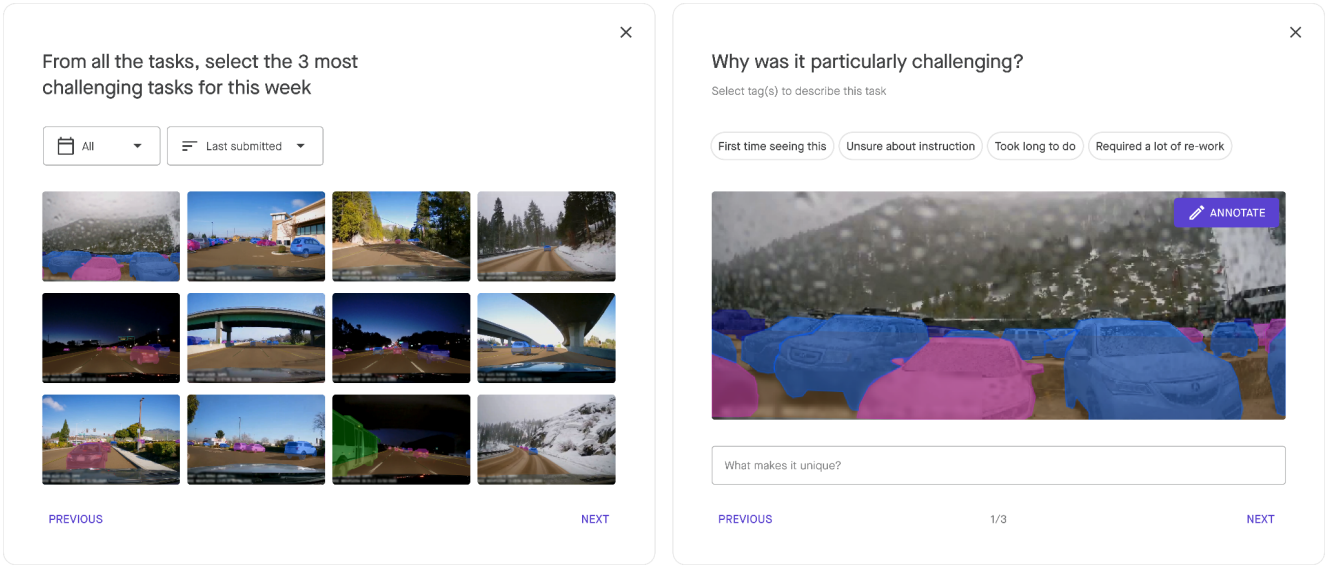


Figure 2: Weekly Reflection, a tool designed to motivate workers to share knowledge they've gained throughout the week.

In response, we present the Calibration Log (Figure 1), an interface devised with the data enrichment worker in mind. Our feedback mechanism is designed to be both intuitive and instructive, facilitating users to easily tag, provide descriptions of challenges faced, and even upload relevant images for clearer context. When faced with uncertainties, workers can momentarily sideline tasks until additional clarity is given. Such flagged tasks are cataloged in the Calibration Log, making them readily accessible for ML professionals' review. This process ensures an immediate line of communication to address potential issues and promotes an atmosphere of transparent and efficient dialogue.

## Weekly Reflection

The annotation industry commonly uses Time Per Task (TPT) to gauge the cognitive effort exerted by the data enrichment workers and the resulting value created (Su, Dang, and Fei-Fei, 2012). Though TPT offers a straightforward metric for various tasks, it is noisy and does not explain the intricacies of why certain tasks might be especially challenging or spotlight critical insights to enhance annotation and validation procedures and outcomes.

To address this gap, we introduce a tool (Figure 2) to allow annotators to spotlight tasks they found most challenging. At the end of the week, this tool prompts them to reflect on their work, offering a gentle and user-friendly interface for introspection, and a comprehensive review of their work. They are presented with a selection of their week's tasks and invited to pinpoint which were most challenging. Moreover, they are encouraged to add descriptive tags, comments, and share any invaluable insights beneficial for the ML engineering team. Our hypothesis is twofold: this optional process provides a platform for annotators to voice their experiences and articulate the dedication they bring to their role. Simultaneously, it is an excellent means to unearth data insights that existing ML models might overlook.

## Work in Progress

The process of co-creating and evaluating these tools is in motion. Version 1 of the Calibration Log interface is planned to be launched in production shortly and Weekly Reflection is being iterated on. Through pilot projects that involve both ML engineering teams from our clients and our data enrichment workforce, we aim to measure any impact the introduction of these collaborative tools might have on their respective projects, and ultimately on the performance of the ML models at stake. In parallel, we are running co-creation workshops to design more of these tools with our data enrichment worker colleagues. Taken together, these initiatives follow our core belief that the next iteration of successful ML model development will be unlocked by providing the people who are closest to the data with means to let their expertise shine.

## Acknowledgement

We thank our colleagues in global service delivery for their ongoing support and openness for this research.

## References


Borkman, S.; Crespi, A.; Dhakad, S.; Ganguly, S.; Hogins, J.; Jhang, Y.C.; Kamalzadeh, M.; Li, B.; Leal, S., Parisi, P.; and Romero, C. 2021. Unity perception: Generate synthetic data for computer vision. arXiv preprint arXiv:2107.04259.

Dai, P.; Rzeszotarski, J.M.; Paritosh, P.; and Chi, E.H. 2015, February. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. doi.org/10.1145/2675133.2675260.

Emam, Z.; Kondrich, A.; Harrison, S.; Lau, F.; Wang, Y.; Kim, A.; and Branson, E. 2021. On the state of data in computer vision: Human annotations remain indispensable for developing deep learning models. arXiv preprint arXiv:2108.00114.

Fernández-Moreno, M.; Lei, B.; Holm, E.A.; Mesejo, P.; and Moreno, R. 2023. Exploring the trade-off between performance and annotation complexity in semantic segmentation. *Engineering Applications of Artificial Intelligence*, 123: 106-299. doi.org/10.1016/j.engappai.2023.106299

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Thieringer, T.; Bethge, M.; Wichmann, F.A.; and Brendel, W., 2021. Partial success in closing the gap between human and machine vision. arXiv:2010.08377

Gray, M. L.; Suri, S.; Ali, S. S.; and Kulkarni, D. 2016. The crowd is a collaborative network. In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing. doi.org/10.1145/2818048.2819942

Irani, L. C.; and Silberman, M. S. . 2013. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1-10. doi.org/10.1145/2470654.2470742

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. Worker motivation in crowdsourcing – a study on mechanical turk. In Proceedings of the Seventeenth Americas Conference on Information Systems. Detroit.

Kittur, A.; Nickerson, J.V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In Proceedings of the 2013 conference on Computer supported cooperative work. doi.org/10.1145/2441776.2441923

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; and Duerig, T. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* 128(7): 1956-1981. doi.org/10.48550/arXiv.1811.00982

Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C.L. 2015. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich. Switzerland. doi.org/10.1007/978-3-319-10602-1_48

Manam, V. K.; and Quinn, A. 2018. Wingit: Efficient refinement of unclear task instructions. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. doi.org/10.1609/hcomp.v6i1.13338

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. Proceedings of the ACM on Human-Computer Interaction. doi.org/10.1145/3415186

Partnership on AI. 2021. Responsible Sourcing of Data Enrichment Services. http://partnershiponai.org/wp-content/uploads/2021/08/PAI-Responsible-Sourcing-of-Data-Enrichment-Services.pdf. Accessed: 2023-07-02

Schilling, M.P.; Rettenberger, L.; Münke, F.; Cui, H.; Popova, A.A.; Levkin, P.A.; Mikut, R.; and Reischl, M. 2021. Label assistant: a workflow for assisted data annotation in image segmentation tasks. In Proceedings—31. Workshop Computational Intelligence. Berlin.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In Workshops at the twenty-sixth AAAI conference on artificial intelligence.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; ... and Anguelov, D. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. doi.org/10.48550/arXiv.1912.04838

Wood, E.; Baltrušaitis, T.; Hewitt, C.; Dziadzio, S.; Cashman, T.J.; and Shotton, J. 2021. Fake it till you make it: face analysis in the wild using synthetic data alone. In Proceedings of the IEEE/CVF international conference on computer vision. doi.org/10.48550/arXiv.2109.15102